Wavelet Subband Filters for Mixed Excitation Linear Predictive (MELP) Vocoder

Naofumi Aoki and Kunio Takaya†

Research Institute for Electronic Science, Hokkaido University

N-12 W-6 Sapporo, 060-0812, Japan. email: naofumi.aoki@ma9.seikyou.ne.jp

†Electrical Engineering, University of Saskatchewan

57 Campus Drive, Saskatoon, Sask. S7N 5A9 , Canada

Phone: (306) 966-5392 Fax: (306) 966-5407 email: kunio_takaya@engr.usask.ca

ABSTRACT

MELP (Mixed Excitation Linear Prediction) vocoder produces much improved voice quality compared with the traditional LPC vocoder. A LPC synthesis filter which mimics the vocal tract characteristics is excited with the mixture of pulse and noise instead of either one alone. This paper describes an implementation of MELP which uses the lifting wavelet transform in place of the bandpass filter bank required for sub-band division in the synthesis and analysis of the MELP vocoder. The mixture ratio between the pulse (voiced) and the noise (unvoiced) is controlled individually in each of the 5 sub-bands considered in this wavelet transform. A new method to generate an appropriate glottal waveform is also described. In addition, three kinds of fluctuations observed in the steady parts of voiced speech are incorporated to enhance the naturalness of synthesized speech. This MELP vocoder was realized on a DSP system based on TMS320C62 to confirm the feasibility of real-time implementation.

I. Introduction

MELP vocoder as a speech codec is more advantageous when the data rate is limited. Its voice quality is potentially comparable to 4.8 kbps CELP (Code Excited Linear Predictive) vocoders, in spite that it runs at 2.4 kbps data rate [6]. The main concept of MELP vocoder is to perform voiced/unvoiced decision in each subband and to generate excitation signals based on the decision. Mixing periodic and aperiodic components in voiced excitation signals contributes to enhance the naturalness of synthesized speech. In the MELP vocoder, usually 5 subbands are considered to analyze their spectral components for the decision making. A filter bank consisting of 5 bandpass filters are used for this purpose. Since the wavelet transform can efficiently divide speech signals into subband components, the filter bank can be replaced by the wavelet transform. A fast algorithm of the wavelet transform, lifting scheme [3], was employed to perform the subband division in our MELP implementation. The lifting scheme is suited particularly for the real-time implementation of MELP vocoder because of its computational efficiency. The phenomena of these fluctuations and their signal models are presented in a later section of this paper.

Another factor which significantly influences the voice quality is the characteristics of glottal waveform [5, 6]. This paper mentions a new method to modify the triangular pulse used in voiced excitation, so that its excitation ratio that determines voice quality can be adjusted. In order to improve the frequency characteristics of a triangular pulse which tends to degrade at high frequencies, the random fractalness observed in source signals obtained by LPC inverse filtering was added to the triangular pulse. In an acoustically clean environment, the MELP vocoder works as a normal LPC vocoder, since every subband tends to decide voiced speech as "voiced" due to the high signal-to-noise ratio. Under such circumstances, the voice tends to be buzzer-like. In order to mitigate this degradation, three kinds of fluctuations which are always observed in the steady parts of voiced speech were incorporated into our MELP vocoder [1].

II. Lifting Wavelet Transform

The lifting scheme is a fast algorithm of the wavelet transform [3]. Low-pass and high-pass filtering by convolution used in the classical wavelet transform procedure are avoided. Instead as shown in Fig. 1, the lifting scheme follows the process of (1) splitting an original sequence into an even and an odd sequence, (2) subtracting the prediction estimated by the even sequence from the odd sequence, and (3) updating the even sequence with the estimate in order to avoid aliasing effects [3]. Splitting the even sequence results in further dividing the signal into two subbands. Applying this procedure make the even sequence represent the coefficients of the scaling function and the odd sequence to represent the wavelet coefficients. The lifting scheme can reduce the computational redundancy involved in the classical scheme. Furthermore, in-place calculation can be performed to save the extra memory needed to store the results of convolution. The scaling function and wavelet used in this study is shown in Fig. 2. The prediction utilizes four adjacent even samples whose weights are (-1/16, 9/16, 9/16, -1/16). This is called cubic interpolating wavelet transform [3]. The lifting wavelet transform of cubic interpolation requires 9 FLOPS (floating operations) for calculating a coefficient of scaling function and a wavelet coefficient, while the classical implementation requires 17 FLOPS. Figure 3 shows the frequency characteristics of the five subbands divided by the lifting wavelet transform. Since the prediction uses only four even samples,

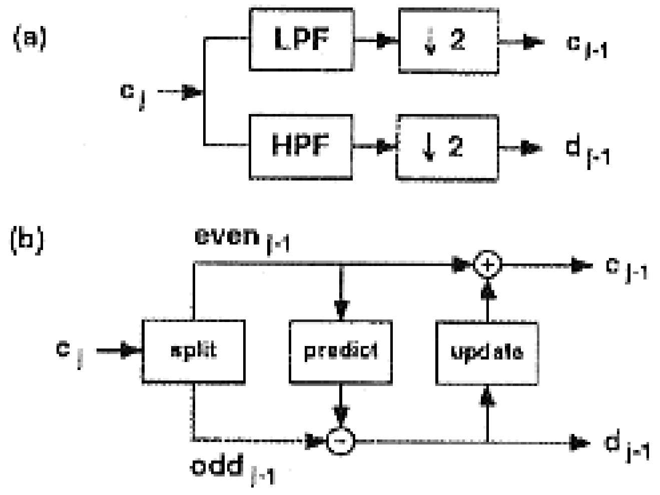

Figure 1: Wavelet transform: (a) classical implementation, (b) lifting scheme.

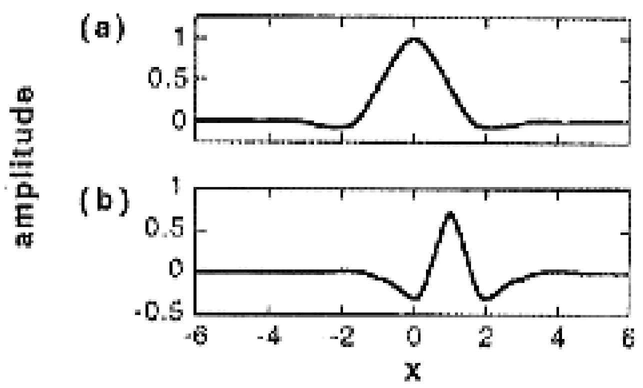

Figure 2: Cubic interpolating wavelet transform: (a) scaling function, (b) wavelet.

subbands are not well separated. This problem can be resolved by increasing the number of samples used in the prediction. However, this leads to increasing the computation by the factor of two [3]. Priority is given to computational efficiency over the precise subband division for the sake of real-time implementation. Figure 4 shows the examples of the reconstructed signal in each subband.

Voiced/unvoiced decision was made in each of the 5 subbands by evaluating the magnitude of normalized autocorrelation function of the wavelet coefficients (level -1 to -4) as well as the coefficients of scaling function (below level -4) allowing a lag of the estimated pitch period [6]. Since the number of the samples available for the auto-correlation calculation is getting smaller progressively at lower levels, this also reduces the computation. In order to enhance the robustness of the decision making, a normalized autocorrelation function after rectifying and smoothing the coefficients was also attempted [6]. Smoothing was performed by a -6 dB/obtave low-pass filter. Pitch period was estimated from the periodicity in the autocorrelation function of the residual signal smoothed by the same -6 dB/octave low-pass filter.

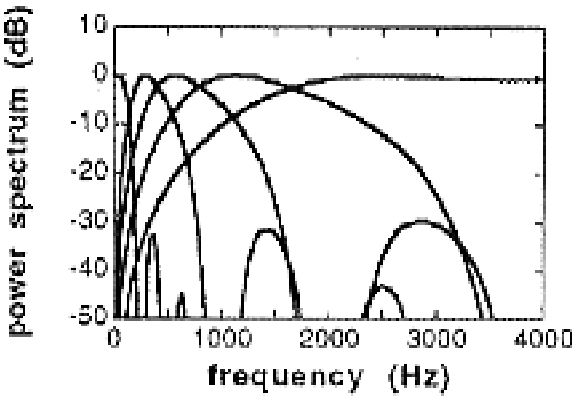

Figure 3: Frequency characteristics of five subbands.

III. Modifying triangular pulse by random fractal interpolation

The signal model of the excitation signals for LPC vocoders are defined as spectral -6 dB/octave in the frequency domain, which includes -12 dB/octave glottal and $+6$ dB/octave radiation characteristics from the mouth [5]. As mentioned in the literature [6], excitation signals that employ a triangular pulse are considered to be more proper for synthesizing human-like natural voice quality, since it has no discontinuities as observed in the conventional rectangular pulse excitation. However, the frequency characteristics of triangular pulses do not meet this required frequency characteristics for **LPC** vocoders, especially in the high frequency region [6].

The graph (a) in the upper panel of Fig. 5 shows an example of triangular pulses for which main excitation ratio (ER) is 32/512, where main excitation ratio is defined as a ratio of the largest to the fastest change in an excitation signal. The frequency characteristic of the triangular pulse is shown by graph (a) in the bottom panel of Fig. 5. Apparently, the frequency characteristic decays faster than -6 dB/oct. In order to mitigate this problem, a technique called random fractal interpolation was devised [2]. It takes advantage of the random fractalness observed in the source signals obtained by the LPC inverse filtering. The curve (b) in the upper panel of Fig. 5 shows an example of such triangular pulses that are modified by the proposed method. The lower panel of Fig. 5 shows that the frequency characteristics of the modified triangular pulse become more compliant with the signal model of LPC vocoders, since it is approximately -6 dB/octave. The modified triangular pulse, however, still maintains almost the same excitation ratio as the original triangular pulse even after the frequency characteristics is changed.

Voice quality of synthesized speech changes as a function of the excitation ratio [4]. The smaller the excitation ratio, the clearer is the voice. The larger the excitation ratio, the softer sounds the voice. In this implementation, the excitation ratio was determined to be inversely

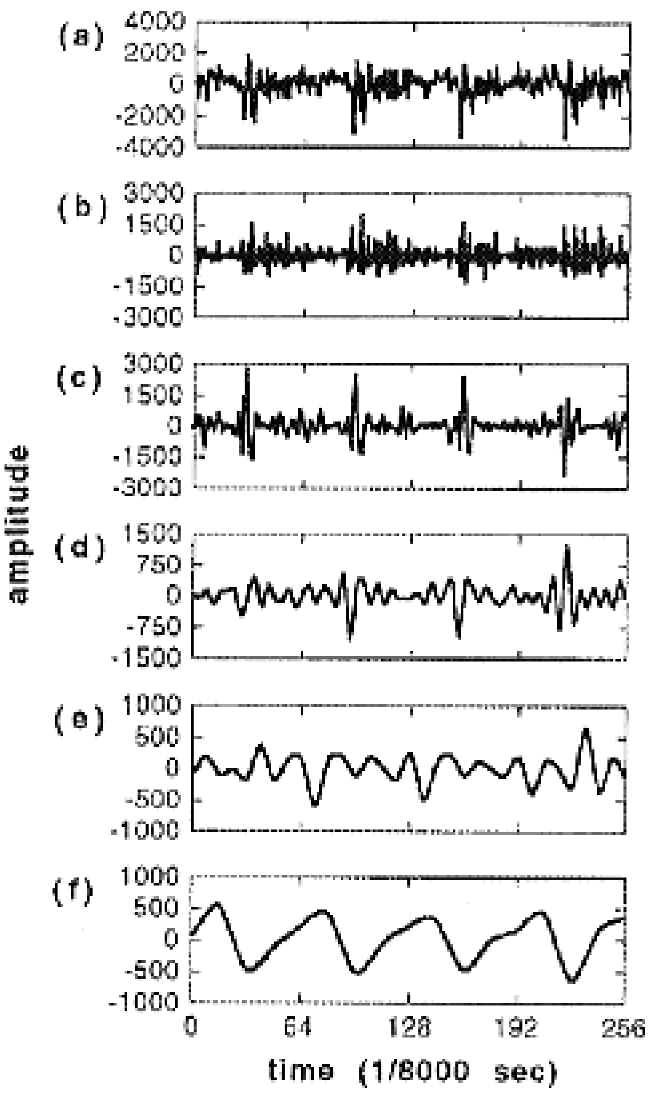

Figure 4: (a) Residual signal. (b) – (f) Reconstructed signals in five subbands: (b) level −1, (c) level −2, (d) level −3, (e) level −4, and (f) less than level −4.

proportional to the peakiness defined in Eq.(1) [4, 6].

$$p = \frac{\frac{1}{N}\sum_{n=0}^{N-1} s_n^2}{(\frac{1}{N}\sum_{n=0}^{N-1} |s_n|)^2} \tag{1}$$

IV. Three types of fluctuations in steady voiced speech

Even in the most steady part of voiced speech, speech signals are not completely periodic. Fluctuations in pitch period and in maximum amplitude are always observed. In addition, waveform itself changes slightly from a pitch period to a pitch period. These three types of fluctuations are considered to be the contributing factors for the naturalness of voiced speech [1]. Our earlier study indicates that pitch period fluctuation and maximum amplitude fluctuation can be modeled as $1/f$ fluctuation [1]. The model for waveform fluctuation can be simply a white noise when excitation signals are regarded as −6 dB/oct [5]. When implementing these fluctuations in the MELP vocoder, the standard deviation for the pitch period was set to 0.05 msec, and the coefficient of variation was set to 7.5 %. The power ratio of the waveform fluctuation to the modified triangular pulse was set to −30 dB. An "aperiodic flag" which indicates the peakiness defined by Eq.(1) [6] was incorporated in the MELP

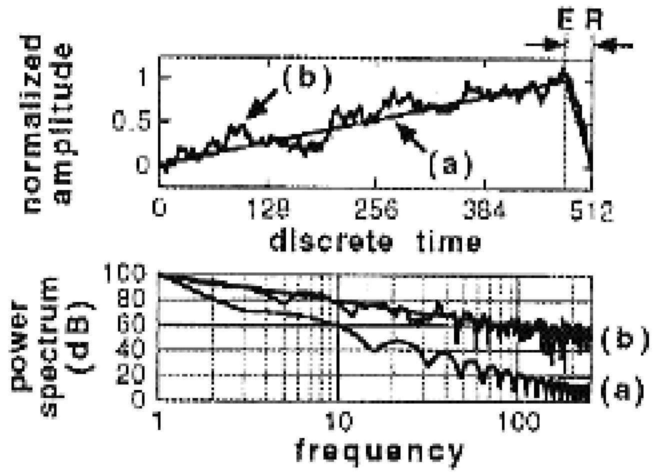

Figure 5: (a) Triangular pulse. (b) Result of random fractal interpolation.

vocoder. When the aperiodic flag was true, the standard deviation of pitch period fluctuation was increased to 0.1 msec.

V. Implementation of MELP vocoder

A block diagram for the synthesis stage of our MELP vocoder is shown in Fig. 6. For the voiced speech, the modified triangular pulse was repeatedly concatenated to generate periodic excitation signals. Then, the wavelet transform was applied to the excitation signals. Using the voiced/unvoiced information for each subband obtained in the analysis stage, random wavelet coefficients were added to such subbands that are marked as unvoiced.

This MELP vocoder was implemented on a DSP evaluation board (TMS320C62 [7]). to confirm a feasibility to execute in real-time the MELP vocoder as described in this paper. The MELP vocoder was programmed in C language. Compared with ordinary LPC vocoders, the voice quality was more natural. Buzzer-like voice was also less noticeable. Three kinds of fluctuations incorporated into our MELP vocoder have significantly enhanced the naturalness. Even when all subbands are switched to the voiced excitation, no buzzer-like voice was heard. Especially, the naturalness of female speech which tends to degrade with the ordinary LPC vocoder was remarkably improved.

VI. Applications

An immediate application of the developed MELP vocoder is to use it as a low bit-rate voice codec. when compared with the conventional LPC vocoder and the 4.8 kbps CELP vocoder commercially used in telephone companies, the voice quality of our developed MELP vocoder is comparable to superior even for voice spoken under noisy environment. The number of bits to be transmitted within a time frame of 25 ms is 60 bits in total, including 7 bits of gain, 7 bits of pitch information, 5 bits of voiced/unvoiced flags (1 bit/band), 40 bits of 10 PARCOR coefficients. This is equivalent to 2,400 bps, which is more compressed than the 4.8 kbps CELP. Technically speaking, a TMS-320C62 200 MHz processor can handle full duplex mode 2.4 kbps MELP at 83% of its com-

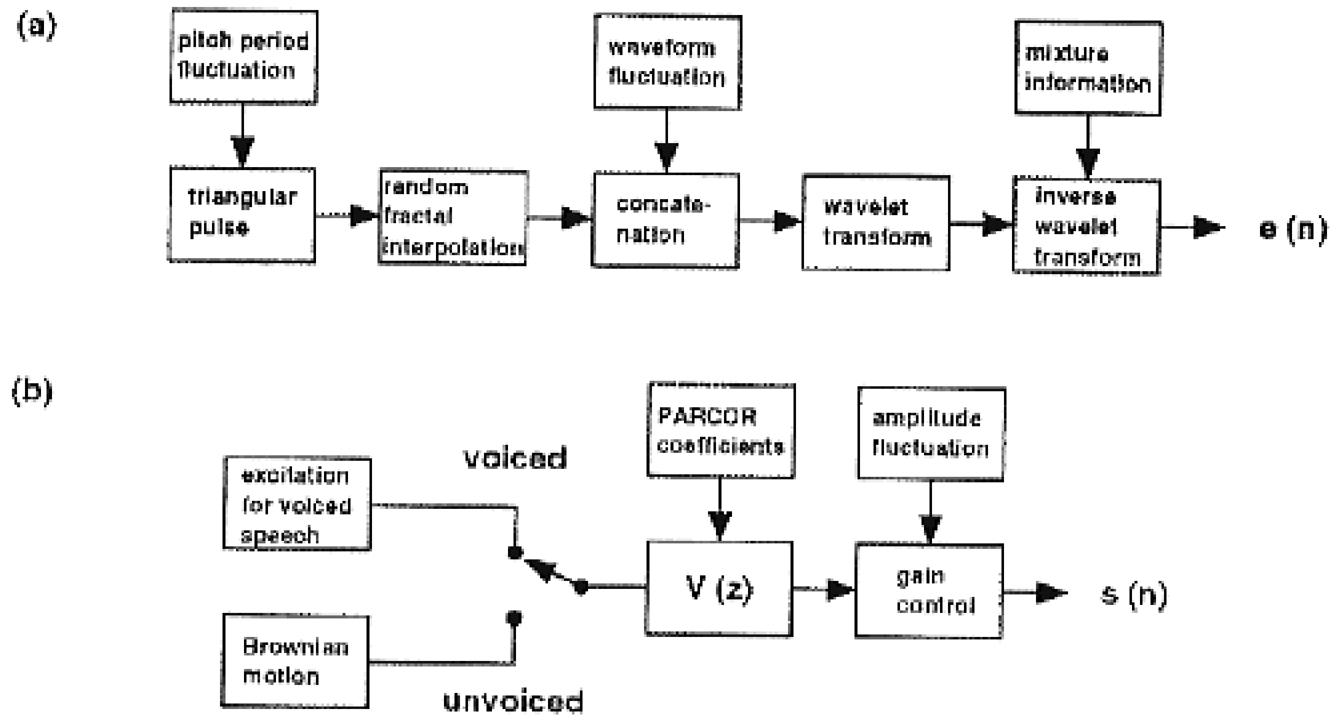

Figure 6: Blockdiagram of the synthesis stage of the MELP vocoder: (a) excitation for voiced speech, (b) speech synthesizer.

putatioal capacity, allowing simultaneous voice analysis (sending) and voice synthesis (receiving).

Another application under development is rule-based speech synthesis which intends to convert a text into speech. The improved naturalness of our MELP vocoder can offer high quality voice for text readers, which is one of the multi-media modalities that enhance user friendliness in the networked computer environment. It has been so far found that MELP vocoder outperforms other vocoders in voiced consonants such as /za/, /ji/, /zu/, /ze/ and /zo/ often used in Japanese language. Since voiced consonants consist of periodic voiced speech and aperiodic unvoiced sound, the MELP vocoder mixing these two kinds of excitation produces far more natural voice.

VII. Concluding Remarks

This study confirmed that mixing periodic and aperiodic components in excitation signals considerably enhanced the naturalness of the synthesized speech. Despite the added features introduced based on the preceding research efforts, i.e. inclusion of the fractal interpolation and three types of fluctuations found in voiced speech, the MELP vocoder was successfully implemented in its entirety by using a TMA320C62 DSP system. The success largely owes the lifting scheme of the wavelet for its computational efficiency in subband filtering.

References

[1] N. Aoki and T. Ifukube, "Fractal modeling of fluctuations in the steady part of sustained vowels for high quality speech synthesis," IEICE Trans. Fundamentals, vol.E81-A, pp.1803–1810, 1998.

[2] N. Aoki and T. Ifukube, "Fractal interpolation for the modification of pulse source signal in PARCOR synthesizer," Int. Conf. Signal Processing, Beijing, pp.650–653, Oct. 1998.

[3] A. Fournier, "Wavelets and their application to computer graphics," Siggraph Course Notes 26, Siggraph, 1995.

[4] T. Hamagami, "Speech synthesis using source wave shape modification technique by harmonic phase control," J. Acoust. Soc. Jpn 54, pp.623–631, 1998 (in Japanese).

[5] D.H. Klatt and L.C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am., vol.87, pp.820–858, 1990.

[6] A.V. McCree and T.P. Barnwell III, "A mixed excitation LPC vocoder model for low bit rate speech coding," IEEE Trans. Speech, Audio Processing, vol.3, pp.242–250, 1995.

[7] Texas Instruments TMS320C62 manuals, http://www.ti.com/sc/c62xevm, 1998.